PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-163140
(43)Date of publication of application : 07.06.2002

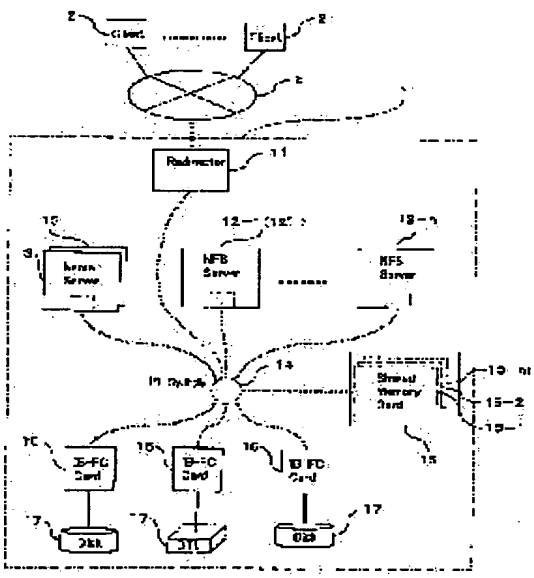
(51)Int.Cl. G06F 12/00
G06F 13/10

(21)Application number : 2000-359810
(22)Date of filing : 27.11.2000

(71)Applicant : FUJITSU LTD
(72)Inventor : OE KAZUICHI
NISHIKAWA KATSUHIKO

(54) STORAGE SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a storage system having a scalability capable of fully coping with the band expansion of a network at a low cost.
SOLUTION: This storage system is provided with a storage device 17 capable of storing file data, a plurality of file servers 12-1 through 12-n performing file processes in response to requests on file data to the storage device 17, a file server management node 11 managing the transfer processes of the file requests received from clients 2 via an external network 3 to the file servers 12-i (i=1 through n) and the response processes to the clients 2 for the file requests, and the internal network 14 communicatably connecting the storage device 17, the file servers 12-i, and the file server management node 11 together.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2002－163140
（P2002－163140A）
(43)公開日　平成14年６月７日(2002.6.7)

(51)Int.Cl.'		識別記号	Ｆ Ｉ		テーマコード'(参考)
Ｇ０６Ｆ	12/00	５４５	Ｇ０６Ｆ 12/00	５４５Ｂ	５Ｂ０１４
	13/10	３４０	13/10	３４０Ａ	５Ｂ０８２

審査請求　未請求　請求項の数５　ＯＬ　（全 16 頁）

(21)出願番号　特願2000－359810(P2000－359810)

(22)出願日　平成12年11月27日(2000.11.27)

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号

(72)発明者　大江　和一
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

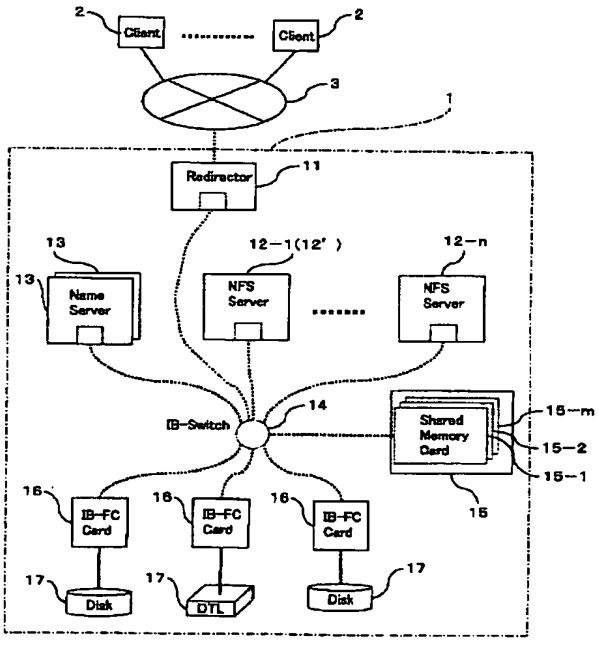
(72)発明者　西川　克彦
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(74)代理人　100092978
弁理士　真田　有

Ｆターム(参考)　5B014 EB04 FA05
5B082 AA01 FA16 HA01 HA05 HA08

(54)【発明の名称】　ストレージシステム

(57)【要約】
【課題】　ネットワークの帯域拡大に対して低コストで十分に対応できるスケーラビリティをもったストレージシステムを提供することを目的とする。
【解決手段】　ファイルデータを記憶しうる記憶装置１７と、ファイルデータに関するリクエストに応じたファイル処理を記憶装置１７に対して行なう複数のファイルサーバ１２－１～１２－ｎと、外部ネットワーク３を介してクライアント２から受信されるファイルリクエストのファイルサーバ１２－ｉ（ｉ＝１～ｎ）への転送処理と、そのファイルリクエストに対するクライアント２への応答処理とを管理するファイルサーバ管理ノード１１と、記憶装置１７，ファイルサーバ１２－ｉ及びファイルサーバ管理ノード１１を通信可能に相互接続する内部ネットワーク１４とをそなえるように構成する。



【特許請求の範囲】
【請求項１】　ファイルデータを記憶しうる記憶装置と、
該ファイルデータに関するリクエストに応じたファイル処理を該記憶装置に対して行なう複数のファイルサーバと、
外部ネットワークを介してクライアントから受信されるリクエストの該ファイルサーバへの転送処理と、該リクエストに対する該クライアントへの応答処理とを一元管理するファイルサーバ管理ノードと、
該記憶装置，該ファイルサーバ及び該ファイルサーバ管理ノードを通信可能に相互接続する内部ネットワークとをそなえて構成されたことを特徴とする、ストレージシステム。
【請求項２】　該内部ネットワークに、該ファイルサーバが扱うファイルデータ名を一元管理するネームサーバが接続されていることを特徴とする、請求項１記載のストレージシステム。
【請求項３】　該内部ネットワークに、該ファイルサーバ管理ノード及び該ファイルサーバがアクセス可能な共有メモリが接続されていることを特徴とする、請求項１記載のストレージシステム。
【請求項４】　該内部ネットワークに、該ファイルサーバ管理ノード，該ファイルサーバ及び該ネームサーバがアクセス可能な共有メモリが接続されていることを特徴とする、請求項２記載のストレージシステム。
【請求項５】　該ファイルサーバ管理ノードが、
該リクエストの内容を解析するリクエスト解析部と、
該リクエスト解析部の解析結果に応じて該リクエストを特定のファイルサーバに転送するリクエスト転送部とをそなえていることを特徴とする、請求項１～４のいずれか１項に記載のストレージシステム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、ストレージシステムに関し、所望のネットワークに接続されることで複数のクライアントでファイルデータの共有を可能にするストレージシステムに関する。
【０００２】
【従来の技術】ネットワーク上での複数ノード（クライアント）間のファイルデータの共有（以下、単に、「ファイル共有」という）を実現する従来の手法としては、例えば図１６に模式的に示すように、ネットワークファイルシステム（ＮＦＳ：Network File System）を利用してＬＡＮ（Local Area Network）などの所望のネットワーク１００上にファイルサーバ２００を構築し、このファイルサーバ２００にＳＣＳＩ（Small Computer System Interface；一般に「スカジー」と呼ばれる）などのインタフェース３００を介して二次記憶装置４００を接続して、この二次記憶装置４００において複数クライアント５００間のファイル共有を実現する方法が良く知られている。
【０００３】しかしながら、この方法では、次のような課題がある。
①ファイルサーバを構築・維持（保守）するのに専門的なスキルが必要である。
②ファイルサーバの拡張（容量，アクセス性能）が容易でない。拡張できてもファイルサーバが複数に分かれたりしてしまうなどで維持コストが増大してしまう。
【０００４】③故障時にそなえたシステム構築・維持（保守）に専門的なスキルが必要であり、また、そのための費用もかかる。
これらの課題を解決する方法として、近年、ＮＡＳ（Network Attached Storage）が提案されている。このＮＡＳは、上記のファイルサーバ２００及び二次記憶装置４００から成る部分（図１６中の破線枠参照）が予め１つのストレージシステムとして構築されたものに相当し、ネットワーク１００に接続して簡単な設定を行なうだけで、ファイル共有が実現できるシステムで、システムの構築・維持（保守）に専門的なスキルは必要ない。
【０００５】
【発明が解決しようとする課題】しかしながら、このようなＮＡＳにおいても、現在急速に進んでいるＬＡＮの帯域拡大（現状で１Ｇｂｐｓ，数年後には１０Ｇｂｐｓ程度）に低コストで十分に対応できるスケーラビリティが得られていないという課題は残っている。即ち、接続先のネットワークの帯域拡大に対応しようとすると、ＮＡＳにおいても、単純に、内部のファイルサーバ及び二次記憶装置を増設することになり、この結果、ファイルサーバが複数に分かれてしまい、それぞれが管理する二次記憶装置も分割されてしまうことになる。
【０００６】つまり、上記のファイルサーバ２００及び二次記憶装置４００を並列（独立）して設けることになる。このため、結局、ファイルサーバの維持（保守）を個々に行なう必要があり、維持コストが増大してしまう。本発明は、このような課題に鑑み創案されたもので、ネットワークの帯域拡大に対して低コストで十分に対応できるスケーラビリティをもったストレージシステムを提供することを目的とする。
【０００７】
【課題を解決するための手段】上記の目的を達成するために、本発明のストレージシステム（請求項１）は、ファイルデータを記憶しうる記憶装置と、この記憶装置に対しリクエストに応じたファイル処理を行なう複数のファイルサーバと、外部ネットワークを介してクライアントから受信されるリクエストの上記ファイルサーバへの転送処理と、そのリクエストに対するクライアントへの応答処理とを一元管理するファイルサーバ管理ノードと、上記の記憶装置，ファイルサーバ及びファイルサーバ管理ノードを通信可能に相互接続する内部ネットワー

クとをそなえて構成されたことを特徴としている。

【０００８】ここで、上記の内部ネットワークには、上記のファイルサーバが扱うファイルデータ名を一元管理するネームサーバが接続されていてもよいし（請求項２）、上記のファイルサーバ管理ノード及びファイルサーバがアクセス可能な共有メモリが接続されていてもよい（請求項３）。なお、上記のネームサーバが存在する場合は、上記のファイルサーバ管理ノード及びファイルサーバに加えて、このネームサーバも上記の共有メモリにアクセスが可能である（請求項４）。

【０００９】また、上記のファイルサーバ管理ノードは、例えば、上記のリクエストの内容を解析するリクエスト解析部と、このリクエスト解析部の解析結果に応じて該リクエストを特定のファイルサーバに転送するリクエスト転送部とをそなえて構成されるのが好ましい（請求項５）。

【００１０】

【発明の実施の形態】以下、図面を参照して本発明の実施の形態を説明する。

（Ａ）一実施形態の説明

図１は本発明の一実施形態としてのストレージシステムの構成（ストレージアーキテクチャ）を示すブロック図で、この図１に示すストレージシステム１（以下、単に「システム１」ともいう）は、外部ネットワーク（例えば、ギガビットイーサネット（登録商標））３に接続された複数のクライアント２間でファイル共有を実現するためのものであって、リダイレクタ（Redirector）１１，複数のＮＦＳサーバ（ファイルサーバ）１２－１～１２－ｎ，ネームサーバ１３，共有メモリ（Shared Memory）１５，ＩＢ－ＦＣカード１６，二次記憶装置１７〔ディスク装置やテープ装置（ＤＴＬ）など〕をそなえており、これらの各コンポーネント１１，１２－ｉ，１３，１５，１６，１７が例えば４～１０Ｇｂｐｓ（ギガビット／秒）程度の高速（内部）ネットワーク〔インフィニバンド（Infiniband）〕スイッチ）１４を介して相互に接続された構成になっている。

【００１１】このようにシステム内部の各コンポーネント（リダイレクタ１１，ＮＦＳサーバ１２－ｉ，ネームサーバ１３，二次記憶装置１７など）を内部ネットワーク１４で接続する形態（クラスタリング）をとることで、必要に応じてＮＦＳサーバ１２－ｉや二次記憶装置１７を内部ネットワーク１４に接続すれば簡単に増設することができ、外部ネットワーク３の通信速度などに応じた拡張性（容量，アクセス性能など）が大幅に向上する。

【００１２】ここで、上記のリダイレクタ（ファイルサーバ管理ノード）１１は、外部ネットワーク３を介して任意のクライアント２から受信される各種リクエストメッセージ（以下、単に「リクエスト」という）のＮＦＳサーバ１２－ｉ（ｉ＝１～ｎ）への転送処理と、そのリクエストに対するリクエスト元のクライアント２への応答処理とを一元（集中）管理するためのものである。つまり、本リダイレクタ１１の存在により、上述のごとくＮＦＳサーバ１２－ｉや二次記憶装置１７を増設したとしても、従来のようにそれらを個々に保守する必要は無いのである。

【００１３】なお、上記の「リクエスト」は、二次記憶装置１７に記憶されたファイルデータ（以下、単に「ファイル」ともいう）に関する要求を意味し、例えば、ファイルデータの実体に対するファイル操作（書き込み／更新／読み出しなど）要求（ファイルアクセスリクエスト）や、それ以外のファイル名参照などのメタ情報アクセスなどがある。また、各クライアント２からは、基本的に、このリダイレクタ１１に付与されたＩＰ（Internet Protocol）アドレスのみが参照できる。つまり、各クライアント２からは、本システム１が１つのファイルサーバとして見えるようになっている。

【００１４】そして、上述したような機能を実現すべく、本リダイレクタ１１には、例えば図２に示すように、外部ネットワーク３用のインタフェースを装備するギガビットイーサネットカード１１ａ、システム内部（内部ネットワーク）用のインタフェースを装備するインフィニバンドカード１１ｂ、これらの各カード１１ａ，１１ｂをはじめとして自身（リダイレクタ１１）の動作を集中制御するためのネットワークプロセッサ１１ｃ、このネットワークプロセッサ１１ｃが動作する上で必要なソフトウェア（プログラム）や各種データを記憶するためのメモリ（主記憶部）１１ｄなどが実装される。

【００１５】なお、ネットワークプロセッサ１１ｃは、これらのコンポーネント１１ａ～１１ｄとＰＣＩ（Peripheral Component Interconnect）バス１１ｅを介して相互通信可能に接続されている。ここで、上記のネットワークプロセッサ１１ｃは、内部ネットワーク１４と外部ネットワーク３との間で送受されるリクエストやそのリクエストに対するリプライ（応答）メッセージなどの送受（プロトコル変換なども含む）や、クライアント２から受信されるリクエスト（プロトコル）の解析，その解析結果に基づいたアクセスファイル名の決定，受信リクエストの転送先ＮＦＳサーバ１２－ｉの決定などを行なうことができるもので、本実施形態では、次のような制御も可能になっている。

【００１６】即ち、クライアント２からのリクエストを解析し、各ファイルサーバ１２－ｉに均一に（例えば、負荷の軽いＮＦＳサーバ１２－ｉに）リクエストが転送されるように制御したり、同一ファイルについてのリクエストは同じＮＦＳサーバ１２－ｉに割り当てて、ＮＦＳサーバ１２－ｉ間でファイルアクセス競合が起きないように制御したりすることができるようになっている。

【００１７】このため、本ネットワークプロセッサ１１

ｃには、その要部の機能に着目すると、次のような機能部が実装されている。
（１）クライアント２からのリクエストの内容を解析するリクエスト解析部１１１としての機能
（２）このリクエスト解析部１１１の解析結果に応じて受信リクエストを特定のＮＦＳサーバ１２－ｉに転送するリクエスト転送部１１２としての機能
（３）リクエスト転送部１１２による過去のリクエストの転送履歴を（例えば、メモリ１１ｄに）記録する転送履歴記録部１１３としての機能
（４）ＮＦＳサーバ負荷監視デーモン（daemon）１１５をバックグラウンドで実行することにより、各ＮＦＳサーバ１２－ｉの負荷状態を定期的に監視する負荷監視部１１４としての機能
これにより、上記のリクエスト転送部１１２は、上記の転送履歴記録部１１３による転送履歴に基づいて同一ファイル名のファイルに対するリクエストを同一ＮＦＳサーバ１２－ｉに転送したり、ＮＦＳサーバ負荷監視デーモン１１５（負荷監視部１１４）による負荷監視結果に基づいて負荷の低いＮＦＳ１２－ｉに受信リクエストを転送したりすることが可能になる。
【００１８】従って、処理速度が大幅に向上するとともに、各ＮＦＳサーバ１２－ｉに均一にリクエストを転送できるので、一部のＮＦＳサーバ１２－ｉに負荷が集中して障害を引き起こしてしまうようなことを確実に防止することができ、信頼性が大幅に向上する。なお、本ネットワークプロセッサ１１ｃは、例えば、メタ情報アクセスに対するリプライメッセージを内部のキャッシュメモリ１１ｆ（メモリ１１ｄでもよい）にキャッシュしておき（図３参照）、クライアント２からメタ情報アクセスについてのリクエストを受けると、まず、キャッシュメモリ１１ｆ（あるいは、メモリ１１ｄ）をチェックして、ヒットすればリプライメッセージを作成してそのまま（ＮＦＳサーバ１２－ｉやネームサーバ１３に転送することなく）クライアント２側に返す（図４参照）ようにすることもできる。
【００１９】この仕組みは、メタ情報のみならずファイルデータについても適用できる。ただし、この場合、全てのファイルデータをキャッシュするようにすると必要なメモリ容量が増大するので、ネットワークプロセッサ１１ｃでアクセス頻度の高いファイルデータを選び出し、そのファイルデータのみをキャッシュメモリ１１ｆ（あるいは、メモリ１１ｄ）にキャッシュするようにした方が良い。
【００２０】つまり、上記のキャッシュメモリ１１ｆ（あるいは、メモリ１１ｄ）は、リクエスト頻度の高い特定のファイルデータについてのクライアント２へのリプライメッセージをキャッシュしておくキャッシュ部としての機能も果たし、この場合のネットワークプロセッサ１１ｃは、新たなリクエストが同じファイルについてのものであると、キャッシュメモリ１１ｆ（あるいは、メモリ１１ｄ）にキャッシュしておいたリプライメッセージをクライアント２へ返す応答部１１６（図２参照）としての機能も有していることになる。
【００２１】このようにして、アクセス頻度の高い情報（メタ情報，ファイルデータ）に関しては、リダイレクタ１１側でキャッシュして、ＮＦＳサーバ１２－ｉにリクエストを転送することなくリダイレクタ１１で応答を返してしまえば、アクセス頻度が高い情報に対するクライアント２への応答速度が大幅に向上し、本システム１としての処理速度及び処理能力が飛躍的に向上することになる。
【００２２】次に、上記のＮＦＳサーバ１２－ｉは、それぞれ、リダイレクタ１１から転送されてきたリクエストに応じたファイル処理（書き込み／更新／読み出しなど）を、内部ネットワーク（内部ネットワークスイッチ）１４経由で二次記憶装置１７にアクセスして実施したり、そのファイル処理結果をリクエスト元のクライアント２への応答として送付するためのリプライメッセージを生成してリダイレクタ１１に送信したりすることができるものである。
【００２３】なお、ＮＦＳサーバ１２－ｉは、いずれも、ハードウェア的には、例えば図５に示すように、ＣＰＵ（Central Processing Unit）１２ａ，メモリ（主記憶部）１２ｂ及び内部ネットワーク１４とのインタフェース（プロトコル変換など）を装備するインタフェースカード（ＩＢ－ＩＦ）１２ｃをそなえて構成され、メモリ１２ｂに記憶されたＮＦＳサーバソフトウェア（プログラム）をＣＰＵ１２ａが読み取って動作することにより、上述したＮＦＳサーバ１２－ｉとしての機能が実現されるようになっている。
【００２４】さて、ここで、これらのＮＦＳサーバ１２－ｉが、それぞれ独自にファイル名を管理すると、同じファイルデータ実体であるにも関わらず異なるＮＦＳサーバ１２－ｉで管理ファイル名が異なったり、逆に、異なるファイルデータ実体であるにも関わらず異なるＮＦＳサーバ１２－ｉで同じ管理ファイル名になったりといった不都合が生じ、ＮＦＳサーバ１２－ｉ間でファイルアクセス競合が生じる可能性がある。
【００２５】このような不都合を解決するのが上記のネームサーバ１３である。つまり、このネームサーバ１３において、ＮＦＳサーバ１２－ｉからのメタ情報アクセスを一元管理することで、全てのＮＦＳサーバ１２－ｉにおける管理ファイル名空間を１つにして、ＮＦＳサーバ１２－ｉ間でのファイルアクセス競合を回避するのである。従って、このネームサーバ１３をそなえることにより、本システム１によるファイル共有の信頼性が大幅に向上することとなる。
【００２６】なお、図１中に示すように、本ネームサーバ１３には、故障などの異常発生（ダウン）にそなえて

現用と予備用とが存在する。また、これらのネームサーバ１３についても、ハードウェア的には、ＮＦＳサーバ１２－ｉと同様の構成（図５参照）、即ち、ＣＰＵ１３ａ，メモリ（主記憶部）１３ｂ及び内部ネットワーク１４とのインタフェースを装備するインタフェースカード１３ｃが実装されており、この場合も、メモリ１３ｂに記憶されたネームサーバソフトウェア（プログラム）をＣＰＵ１３ａが読み取って動作することにより、上述したネームサーバ１３としての機能が実現されるようになっている。

【００２７】次に、上記の共有メモリ１５は、上記のリダイレクタ１１，ＮＦＳサーバ１２－ｉ及びネームサーバ１３がそれぞれ内部ネットワーク１４を介してアクセス可能なメモリで、例えば、或るＮＦＳサーバ１２－ｉあるいは現用のネームサーバ１３がダウンした場合（障害発生時）に、そのＮＦＳサーバ１２－ｉあるいは現用のネームサーバ１３の処理を他のＮＦＳサーバ１２－ｋ（ｋ＝１～ｎでｋ≠ｉ）あるいは予備用のネームサーバ１３に引継ぐための情報（以下、引継ぎ情報）などがサーバ１２－ｉ，１３別にメモリカード（Shared Memory Card）１５－１～１５－ｍ（ｍは自然数）に保持（バックアップ）されるようになっている（図６及び図８参照）。

【００２８】つまり、上記の各ＮＦＳサーバ１２－ｉ（ＣＰＵ１２ａ）もしくはネームサーバ１３（ＣＰＵ１３ａ）は、異常発生時にそなえて他のＮＦＳサーバ１２－ｉもしくは予備用のネームサーバ１３に処理を引継ぐのに必要な情報を引継ぎ情報として共有メモリ１５に記録する引継ぎ情報記録部１２１（１３１）（図５参照）としての機能を有していることになる。

【００２９】なお、ＮＦＳサーバ１２－ｉのダウンは、例えば図６に模式的に示すように、現用のネームサーバ１３（ＣＰＵ１３ａ）がＮＦＳサーバ監視デーモン１３２をバックグラウンドで実行することで監視し、現用のネームサーバ１３のダウンは、例えば図８に模式的に示すように、予備用のネームサーバ１３（ＣＰＵ１３ａ）がネームサーバ監視デーモン１３３をバックグラウンドで実行することで監視する。

【００３０】そして、図７に模式的に示すように、ＮＦＳサーバ１２－ｉのダウンが検出された場合（ステップＳ１）は、現用のネームサーバ１３（ＣＰＵ１３ａ）が、ダウンしたＮＦＳサーバ１２－ｉ以外のＮＦＳサーバ１２－ｋ（例えば、負荷の軽いＮＦＳサーバ１２－ｋ）に対して、ダウンしたＮＦＳサーバ１２－ｉの処理を引継ぐよう指示するとともに、リダイレクタ１１に対してＮＦＳサーバ１２－ｉのダウンを通知する（ステップＳ２）。

【００３１】これにより、引継ぎ指示を受けたＮＦＳサーバ１２－ｋ（ＣＰＵ１２ａ）は、共有メモリ１５に内部ネットワーク１４を介してアクセスして、そこにバックアップされている引継ぎ情報を読み出してダウンしたＮＦＳサーバ１２－ｉの処理を引継ぐ（ステップＳ３）。一方、このとき、リダイレクタ１１（ネットワークプロセッサ１１ｃ）は、ネームサーバ１３から上記の通知を受けることにより、ダウンしたＮＦＳサーバ１２－ｉにはリクエスト転送部１１２によるリクエストの転送を行なわないようにする。

【００３２】つまり、この場合のネームサーバ１３（ＣＰＵ１３ａ）は、ＮＦＳサーバ１２－ｉの異常発生を検出する異常検出部１３４（図６参照）と、この異常検出部１３４でＮＦＳサーバ１２－ｉの異常発生が検出されると、そのＮＦＳサーバ１２－ｉ以外の他のＮＦＳサーバ１２－ｋに対して異常発生したＮＦＳサーバ１２－ｉの処理を共有メモリ１５の引継ぎ情報に基づいて引継ぐよう引継ぎ指示を与える引継ぎ指示部１３５（図６参照）としての機能を有していることになる。

【００３３】このように、本実施形態では、たとえ、ＮＦＳサーバ１２－ｉやネームサーバ１３がダウンしたとしても、他のＮＦＳサーバ１２－ｋや予備用のネームサーバ１３が処理を引継ぐことができるので、ストレージシステム１としては正常なファイル処理を継続することができ、耐障害性が大幅に向上する。なお、本例は、ＮＦＳサーバ１２－ｉやネームサーバ１３の冗長化についての説明であるが、勿論、リダイレクタ１１を同様にして冗長化することも可能である。また、現用のネームサーバ１３ダウン時の引継ぎについては、場合によってはＮＦＳサーバ１２－ｉのいずれかに引継ぐようにしてもよい。

【００３４】次に、上述したリダイレクタ１１からＮＦＳサーバ１２－ｉへのリクエスト転送時のリダイレクタ１１及びＮＦＳサーバ１２－ｉでの具体的な処理について説明する。リダイレクタ１１は、例えば、クライアント２から或るファイルデータを書き込むためのファイルアクセスリクエストを受信すると、そのファイルアクセスリクエストをリクエスト解析部１１１にて解析する。

【００３５】なお、上記の「ファイルアクセスリクエスト」は、例えば図１２に示すように、物理レイヤヘッダ（Phy Header）２１ａやＩＰヘッダ（Internet Protocol Header）２１ｂ，ＴＣＰヘッダ（Transmission Control Protocol Header）２１ｃ，ＮＦＳヘッダ２１ｄなどから成るヘッダ部２１と、実際に二次記憶装置１７に書き込むべきファイルデータの実体（実ファイルデータ）が格納された実ファイルデータ部２２とを有して成る。

【００３６】そして、リクエスト解析部１１１は、図９に模式的に示すように、上記のファイルアクセスリクエスト中の実ファイルデータが始まる位置、即ち、ヘッダ部２１と実ファイルデータ部２２との境界をヘッダオフセット値〔境界情報；例えば先頭からのビット数（“ａ”）など〕２３として求める。求められた境界情報２３は、リクエスト転送部１１２に通知され、リクエ

スト転送部１１２は、上記ファイルアクセスリクエストに通知された境界情報２３を添付（付加）して転送先のＮＦＳサーバ１２－ｉに送付する。

【００３７】つまり、上記のリクエスト解析部１１１は、図２中に示すように、受信したファイルアクセスリクエストを解析してそのファイルアクセスリクエストのヘッダ部２１と実ファイルデータ部２２との境界位置を表わすヘッダオフセット値２３を求めるヘッダオフセット値解析部１１１ａとしての機能と、このヘッダオフセット値解析部１１１ａで得られたヘッダオフセット値２３をＮＦＳサーバ１２－ｉへ転送されるファイルアクセスリクエストに付加するヘッダオフセット値付加部１１１ｂとしての機能とを有していることになる。

【００３８】その後、ＮＦＳサーバ１２－ｉ側では、上述のごとくリダイレクタ１１側で付加されたヘッダオフセット値２３に基づいて、ＮＩＣ（Network Interface Card）ドライバ（ネットワークドライバ）１２２が、実ファイルデータ部２２とそれ以外の領域（ヘッダ部２１）の先頭アドレスをそれぞれ上位層（ＮＦＳ処理層）のカーネル内（カーネル上位層）で扱われるメッセージのページ境界〔ページ境界（別領域）；バッファ（mbuf）１２３，１２４〕に割り当てることができる（図１０参照）。

【００３９】このようにすると、例えば図１１に模式的に示すように、ファイルアクセスリクエストがカーネル上位層のファイルシステム部１２５に到達したときに、実ファイルデータ部２２の先頭アドレス（ポインタ）をファイルシステムバッファ１２６へのポインタに付け替えるだけで、データのコピーを発生させずに（マップ切り替え）、ファイルシステムバッファ１２６にデータを移すことが可能になる（カーネル内ゼロコピーが実現される）。従って、ＤＭＡ（Direct Memory Access）も高速に実行でき、ＮＦＳサーバ１２－ｉの処理速度及び処理能力を大幅に向上することができる。

【００４０】なお、ヘッダ部２１と実ファイルデータ部２２との境界は、ＮＩＣドライバ１２２側で求めるようにすることも考えられるが、この場合、ＮＩＣドライバ１２２の処理（ヘッダ解析）量が増大する（通常、ＮＩＣドライバ１２２は物理レイヤヘッダ２１ａの解析のみを行なう）ため、上記のように元々ヘッダ部２１の解析機能（リクエスト解析部１１１）を有するリダイレクタ１１側で境界を求めるようにした方が、ＮＩＣドライバ層での処理量を増やさずに（処理能力低下を招かずに）上位層（ＮＦＳ処理層）でのカーネルゼロコピーを実現できる。

【００４１】以上のように、本実施形態のストレージシステム１によれば、システム１内部に、リダイレクタ１１，複数のＮＦＳサーバ１２－ｉ，ネームサーバ１３，共有メモリ１５，二次記憶装置１７を設け、これらを高速な内部ネットワーク１４で接続する構成にしたことにより、必要に応じてＮＦＳサーバ１２－ｉや二次記憶装置１７を簡単に増設することができ、しかも、ＮＦＳサーバ１２－ｉの維持（保守）を個々に行なう必要もないので、外部ネットワーク３の帯域拡大に対して低コストで十分に対応できる（例えば、１０ＧｂｐｓＬＡＮまで対応可能な）性能・容量スケーラビリティを確保することができる。

【００４２】特に、上述した実施形態では、リダイレクタ１１が、各ＮＦＳサーバ１２－ｉの負荷に応じて各ＮＦＳサーバ１２－ｉに均一に処理が割り当てられるように制御したり、同じファイルについてのリクエストに対する処理は同じＮＦＳ１２－ｉに割り当てたり、アクセス頻度の高いファイルについてのリクエストに対してはＮＦＳサーバ１２－ｉ側ではなくリダイレクタ１１側でキャッシュしておいたリプライメッセージにより応答したりするので、その処理速度及び性能が飛躍的に向上しており、１０ＧｂｐｓＬＡＮまで確実に対応可能な性能・容量スケーラビリティが実現されている。

【００４３】（Ｂ）第１変形例の説明

上述したシステム１内には、メモリ１２ｂの容量が他のＮＦＳサーバ１２－ｉよりも大きいＮＦＳサーバ１２－ｉをキャッシュサーバ１２′（図１参照）として配置して、このキャッシュサーバ１２′でのファイルアクセスは基本的にメモリ１２ｂに対する読み書きのみでクライアント２側に応答を返すようにしてもよい。

【００４４】そして、一定期間内のリクエスト頻度が所定回数（閾値）以上であるアクセス頻度の高いファイルについてはキャッシュサーバ１２′のメモリ１２ｂにキャッシュしておき、キャッシュサーバ１２′が応答するようにする。具体的には、まず、リダイレクタ１１側で各ファイルのアクセス頻度を監視しておき、アクセス頻度が或る閾値よりも高いファイルへのアクセスに関しては、キャッシュサーバ１２－ｉで処理を行なうように、リダイレクタ１２－ｉからネームサーバ１３，ＮＦＳサーバ１２－ｉ，キャッシュサーバ１２′に指示を与える。

【００４５】つまり、このとき、リダイレクタ１１（リクエスト転送部１１２）は、アクセス頻度の高いファイルについてのリクエストをキャッシュサーバ１２′に転送することになる。これにより、アクセス頻度の高いファイルに対するアクセスは、二次記憶装置１７にアクセスすることなくキャッシュサーバ１２′内で処理されるので、ストレージシステム１としての処理速度及び処理能力の大幅な向上に大きく寄与する。

【００４６】一方、上記のアクセス頻度の高いファイルに対するアクセス頻度が落ちてくると（キャッシュサーバ１２′のメモリ１２ｂにキャッシュされているファイルに対するアクセス頻度が所定回数以下になると）、リダイレクタ１１が適当な（例えば、負荷の軽い）ＮＦＳサーバ１２－ｉを割り当てて処理を移行するようにネー

ムサーバ１３，ＮＦＳサーバ１２－ｉ，キャッシュサーバ１２′に指示する。

【００４７】つまり、この場合、リダイレクタ１１（リクエスト転送部１１２）は、リクエストの転送先をキャッシュサーバ１２′以外のＮＦＳサーバ１２－ｉに変更するのである。これにより、アクセス頻度が落ちてきたファイルがいつまでもキャッシュサーバ１２′にキャッシュされ続けることが回避され、その結果、キャッシュサーバ１２′に必要なメモリ容量を削減することができるとともに、キャッシュサーバ１２′での処理に余裕をもたせてその処理能力を向上することができる。

【００４８】（Ｃ）第２変形例の説明

なお、例えば図１３に示すように、上記の二次記憶装置１７をＦＣスイッチ１８経由でネームサーバ１３及びＮＦＳサーバ１２－ｉと接続（二次記憶ネットワークを構築）し、ＦＣスイッチ１８と外部ノード１９とを接続することにより、二次記憶装置１７に外部ノード１９からもアクセスできるようにすることも可能である。ただし、この場合、外部ノード１９で動作させるファイルシステムは、ストレージシステム１内のファイルシステムと同じもの（上述した例では、ＮＦＳ；図１３中の網かけ部がこれを意味する）である必要がある。

【００４９】このようにすることで、システム１内部のＮＦＳサーバ１２－ｉとのアクセス競合を避けるために外部ノード１９からのアクセスについては何らかの調停制御が必要になるが、外部ノード１９から本ストレージシステム１内のファイルへのアクセスが可能になる。ただし、例えば図１４に示すように、外部ノード１９からネームサーバ１３へのアクセスを許容するようにすれば、外部ノード１９からのアクセスについてもシステム１内の管理ファイル名に従うことになるので、上記の調停制御を必要とせずに外部ノード１９からのファイルアクセスが行なえるようになる。なお、この図１４では、内部ネットワーク１４経由でのネームサーバ１３へのアクセスを許容する場合であるが、勿論、図１３において二次記憶ネットワーク（ＦＣスイッチ１８）経由でのアクセスを許容するようにしてもよい。

【００５０】また、例えば図１５に示すように、ＮＦＳサーバ１２－ｉと外部ノード１９とを共通化してしまってもよい。つまり、ＮＦＳサーバ１２－ｉを、外部ネットワーク３から直接受信されるリクエストに応じたファイル処理を二次記憶装置１７に対して行なうように構成するのである。これにより、或るクライアント２がリダイレクタ１１経由でアクセスしてきた場合は、ＮＦＳサーバ１２－ｉは上述したストレージシステム１のファイルサーバとして機能し、直接、ＮＦＳサーバ１２－ｉにアクセスしてきた場合は、リダイレクタ１１を経由せずに応答する通常のファイルサーバとして機能することになる。つまり、クライアント２からのＮＦＳサーバ１２－ｉ経由でのアクセスとＮＦＳサーバ１２－ｉを経由しない直接アクセスとの双方を許容するのである。

【００５１】いずれの場合も、外部からの直接アクセスが可能となり、他のストレージアーキテクチャ〔例えば、ＳＡＮ（Storage Area Network）など〕との融合を実現することができる。なお、図１４及び図１５において、符号２０は内部ネットワーク１４と二次記憶装置１７とのインタフェースを装備するネットワークディスクアダプタを表す。

【００５２】また、図１３～図１５に示す構成では、前述した共有メモリ１５が省略されているが、勿論、装備されていてもよい。このようにすれば、図１３～図１５に示す構成においても、前記と同様のバックアップ処理が可能になる。

（Ｄ）その他

なお、上述した実施形態では、内部ネットワーク１４としてInfiniband，外部ネットワーク３としてギガビットイーサネットを適用した場合について説明したが、勿論、これら以外の他の高速ネットワークを用いてシステム構築することも可能である。

【００５３】また、上記のネームサーバ１３や共有メモリ１５は必ずしもそなえる必要はなく、これらのいずれか、あるいは、双方を省略しても本発明の目的は十分達成される。さらに、上述した実施形態では、ファイルサーバにＮＦＳを適用しているが、本発明はこれに限定されず、勿論、他のファイルシステムを適用することも可能である。

【００５４】また、上述した実施形態では、内部ネットワーク１４の容量（通信速度）が４～１０Ｇｂｐｓ程度であることを前提としたが、この速度は外部ネットワーク３の帯域拡大に応じて適宜変更すれば対応できる。そして、本発明は、上述した実施形態に限定されず、上記以外にも本発明の趣旨を逸脱しない範囲で種々変形して実施することができる。

【００５５】（Ｅ）付記

〔付記１〕　ファイルデータを記憶しうる記憶装置と、リクエストに応じたファイル処理を該記憶装置に対して行なう複数のファイルサーバと、外部ネットワークを介してクライアントから受信されるリクエストの該ファイルサーバへの転送処理と、該リクエストに対する該クライアントへの応答処理とを一元管理するファイルサーバ管理ノードと、該記憶装置，該ファイルサーバ及び該ファイルサーバ管理ノードを通信可能に相互接続する内部ネットワークとをそなえて構成されたことを特徴とする、ストレージシステム。

【００５６】〔付記２〕　該内部ネットワークに、該ファイルサーバが扱うファイルデータ名を一元管理するネームサーバが接続されていることを特徴とする、付記１記載のストレージシステム。

〔付記３〕　該内部ネットワークに、該ファイルサーバ管理ノード及び該ファイルサーバがアクセス可能な共有

メモリが接続されていることを特徴とする、付記１記載のストレージシステム。

【００５７】〔付記４〕　該内部ネットワークに、該ファイルサーバ管理ノード，該ファイルサーバ及び該ネームサーバがアクセス可能な共有メモリが接続されていることを特徴とする、付記２記載のストレージシステム。

〔付記５〕　該ファイルサーバ管理ノードが、該リクエストの内容を解析するリクエスト解析部と、該リクエスト解析部の解析結果に応じて該リクエストを特定のファイルサーバに転送するリクエスト転送部とをそなえていることを特徴とする、付記１～４のいずれか１項に記載のストレージシステム。

【００５８】〔付記６〕　該ファイルサーバ管理ノードが、該リクエスト転送部による過去のリクエストの転送履歴を記録する転送履歴記録部をそなえ、該リクエスト転送部が、該転送履歴記録部の該転送履歴に基づいて同一ファイルデータ名のファイルデータに対するリクエストを同一ファイルサーバに転送するように構成されたことを特徴とする、付記５記載のストレージシステム。

【００５９】〔付記７〕　該ファイルサーバ管理ノードが、該ファイルサーバの負荷を監視する負荷監視部をそなえ、該リクエスト転送部が、該負荷監視部での監視結果に基づいて負荷の低いファイルサーバに該リクエストを転送するように構成されたことを特徴とする、付記５記載のストレージシステム。

【００６０】〔付記８〕　該ファイルサーバのうち少なくとも１台が、該記憶装置におけるファイルデータをキャッシュしうる主記憶部をそなえ、該主記憶部において該リクエストに応じたファイル処理を実行するキャッシュサーバとして構成されていることを特徴とする、付記５記載のストレージシステム。

〔付記９〕　該キャッシュサーバの該主記憶部が、一定期間内のリクエスト頻度が所定回数以上のファイルデータをキャッシュしておくように構成されるとともに、該リクエスト転送部が、上記のリクエスト頻度の高いファイルデータについてのリクエストを該キャッシュサーバに転送するように構成されたことを特徴とする、付記８記載のストレージシステム。

【００６１】〔付記１０〕　該リクエスト転送部が、該キャッシュサーバの該主記憶部にキャッシュされている該ファイルデータに対するリクエスト頻度が所定回数以下になると、該リクエストの転送先を該キャッシュサーバ以外のファイルサーバに変更するように構成されたことを特徴とする、付記９記載のストレージシステム。

【００６２】〔付記１１〕　該リクエスト解析部が、該リクエストを解析して当該リクエストのヘッダ部と実ファイルデータ部との境界位置を表わすヘッダオフセット値を求めるヘッダオフセット値解析部と、該ヘッダオフセット値解析部で得られた該ヘッダオフセット値を該ファイルサーバへ転送される該リクエストに付加するヘッダオフセット値付加部とをそなえていることを特徴とする、付記５記載のストレージシステム。

【００６３】〔付記１２〕　該ファイルサーバが、該リクエストに付加された該ヘッダオフセット値に基づいて該リクエストの該ヘッダ部と該データ部とをそれぞれカーネル上位層で扱われるメッセージの異なる領域にコピーするネットワークドライバをそなえていることを特徴とする、付記１１記載のストレージシステム。

〔付記１３〕　該ファイルサーバ管理ノードが、リクエスト頻度の高い特定のファイルデータについての該クライアントへの応答メッセージをキャッシュしておくキャッシュ部と、該リクエストが該特定のファイルデータについてのものであると、該キャッシュ部の該当応答メッセージを該クライアントへ返す応答部とをそなえていることを特徴とする、付記１～１２のいずれか１項に記載のストレージシステム。

【００６４】〔付記１４〕　該ファイルサーバが、異常発生時にそなえて他のファイルサーバに処理を引継ぐのに必要な情報を引継ぎ情報として該共有メモリに記録する引継ぎ情報記録部をそなえていることを特徴とする、付記３又は付記４に記載のストレージシステム。

〔付記１５〕　該ファイルサーバの異常発生を検出する異常検出部と、該異常検出部で該ファイルサーバの異常発生が検出されると、当該ファイルサーバ（以下、異常ファイルサーバという）以外の他のファイルサーバに対して該異常ファイルサーバの処理を該共有メモリの該引継ぎ情報に基づいて引継ぐよう引継ぎ指示を与える引継ぎ指示部とが設けられたことを特徴とする、付記１４記載のストレージシステム。

【００６５】〔付記１６〕　該記憶装置が、外部ノードからのアクセスを許容するように構成されたことを特徴とする、付記１～１５のいずれか１項に記載のストレージシステム。

〔付記１７〕　該ネームサーバが、外部ノードからのアクセスを許容するように構成されたことを特徴とする、付記２～１５のいずれか１項に記載のストレージシステム。

【００６６】〔付記１８〕　該ファイルサーバが、該外部ネットワークから直接受信されるリクエストに応じたファイル処理を該記憶装置に対して行なうように構成されたことを特徴とする、付記１～１７のいずれか１項に記載のストレージシステム。

【００６７】

【発明の効果】以上詳述したように、本発明のストレージシステムによれば、リクエストに応じたファイル処理を記憶装置に対して行なう複数のファイルサーバと、各ファイルサーバの処理を一元管理するファイルサーバ管理ノードと、記憶装置，ファイルサーバ及びファイルサーバ管理ノードを通信可能に相互接続する内部ネットワークとをそなえているので、必要に応じてファイルサー

バや記憶装置を簡単に増設することができ、しかも、各ファイルサーバの維持（保守）を個々に行なう必要もないので、外部ネットワークの帯域拡大に対して低コストで十分に対応できる性能・容量スケーラビリティを確保することができる。

【００６８】そして、上記の内部ネットワークに、上記の各ファイルサーバが扱うファイルデータ名を一元管理するネームサーバを接続すれば、ファイルサーバ間でファイルアクセス競合が生じることを防止することができるので、ファイル共有の信頼性向上に大きく寄与する。また、上記の内部ネットワークに、共有メモリを接続して、この共有メモリに障害発生時にそなえたファイルサーバの引継ぎ情報を随時記憶するようにしておくことで、一部のファイルサーバに障害が発生したとしても、ストレージシステムとしては正常な処理を継続することができるので、耐障害性が大幅に向上する。

【００６９】さらに、ファイルサーバ管理ノードは、過去のリクエストの転送履歴に基づいて同一ファイルデータ名のファイルデータに対するリクエストを同一ファイルサーバに転送するように構成してもよく、このようにすれば、処理速度が大幅に向上する。また、ファイルサーバの負荷を監視して負荷の低いファイルサーバにリクエストを転送するようにしてもよく、このようにすれば、各ファイルサーバに均一にリクエストを転送できるので、一部のファイルサーバに負荷が集中して障害を引き起こしてしまうようなことを確実に防止することができ、信頼性が大幅に向上する。

【００７０】さらに、リクエスト頻度が高いファイルデータについては、キャッシュサーバの主記憶部にキャッシュしておき、キャッシュサーバで処理するようにしておけば、記憶装置へのアクセス頻度を大幅に削減することができるので、さらに処理速度及び処理性能が向上する。そして、この場合、キャッシュサーバの主記憶部にキャッシュされているファイルデータに対するリクエスト頻度が所定回数以下になると、キャッシュサーバ以外のファイルサーバで処理を行なうようにすれば、リクエスト頻度の低いファイルデータがいつまでもキャッシュサーバに保持され続けることが回避されるので、キャッシュサーバの主記憶部に必要なメモリ容量を削減することができるとともに、キャッシュサーバでの処理に余裕をもたせてその処理能力を向上することができる。

【００７１】また、ファイルサーバ管理ノードにおいて、リクエストのヘッダ部と実ファイルデータ部との境界位置を表わすヘッダオフセット値を求めて、そのヘッダオフセット値をリクエストに付加してファイルサーバに転送するようにすれば、ファイルサーバのネットワークドライバにおいて、そのヘッダオフセット値に基づいてリクエストのヘッダ部とデータ部とをそれぞれカーネル上位層で扱われるメッセージの異なる領域にコピーすることができる。従って、カーネル内ゼロコピーを実現することができ、ファイルサーバの処理速度及び処理能力を大幅に向上することができる。

【００７２】さらに、上記のファイルサーバ管理ノードにおいて、リクエスト頻度の高い特定のファイルデータについてのクライアントへの応答メッセージをキャッシュしておき、受信したリクエストがそのファイルデータについてのものであると、キャッシュしておいた応答メッセージをクライアントへ返すようにすれば、ファイルサーバへリクエストを転送する必要が無いので、クライアントに対する応答速度が大幅に向上し、本システムとしての処理速度及び処理能力が飛躍的に向上することになる。

【００７３】また、上記の記憶装置は、外部ノードからのアクセスを許容するように構成してもよく、このようにすれば、他のストレージアーキテクチャとの融合が可能になる。ここで、外部ノードから上記のネームサーバへのアクセスを許容するようにすれば、上記のファイルサーバと外部ノードとのファイルアクセス調停制御を必要とせずに、外部ノードからのファイルアクセスが可能になる。

【００７４】さらに、上記のファイルサーバは、上記の外部ネットワークから直接受信されるリクエストに応じたファイル処理を記憶装置に対して行なうように構成してもよく、このようにすれば、クライアントからのファイルサーバ経由でのアクセスとファイルサーバを経由しない直接アクセスとの双方を許容することができるので、この場合も、他のストレージアーキテクチャとの融合が可能になる。

【図面の簡単な説明】

【図１】本発明の一実施形態としてのストレージシステムの構成（ストレージアーキテクチャ）を示すブロック図である。

【図２】図１に示すリダイレクタの構成を示すブロック図である。

【図３】図１に示すリダイレクタでメタ情報をキャッシュする場合を説明するためのブロック図である。

【図４】図１に示すリダイレクタでリプライメッセージを返す場合を説明するためのブロック図である。

【図５】図１に示すＮＦＳサーバ（ネームサーバ）の構成を示すブロック図である。

【図６】図１に示す共有メモリにＮＦＳサーバの引継ぎ情報をバックアップする場合を説明するためのブロック図である。

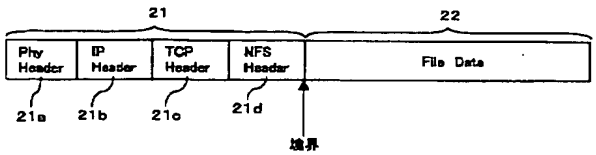
【図７】図６に示す共有メモリにバックアップされた引継ぎ情報に基づいてダウンしたＮＦＳサーバの処理を引継ぐ場合を説明するためのブロック図である。

【図８】図１に示す共有メモリにネームサーバの引継ぎ情報をバックアップする場合を説明するためのブロック図である。

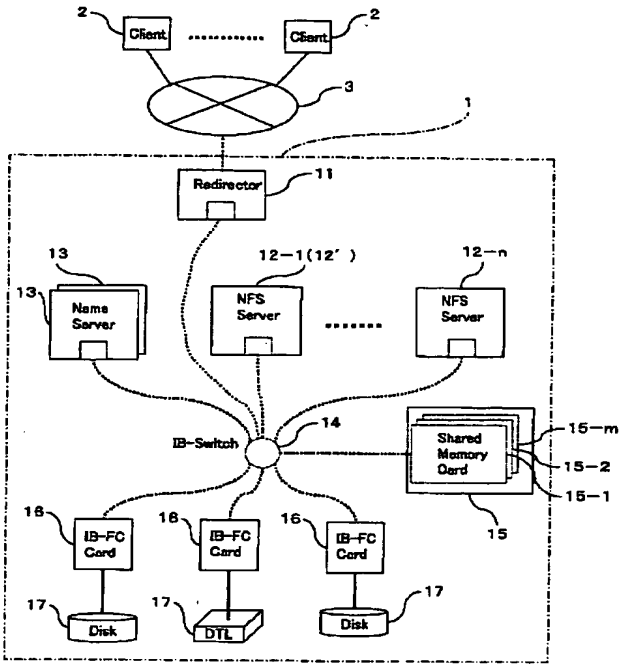
【図９】図１に示すリダイレクタにおいてファイルアク

セスリクエストの境界情報を求める場合を説明するためのブロック図である。
【図１０】図１に示すＮＦＳサーバにおいて図９に示すファイルアクセスリクエストの境界情報に基づいてカーネル内ゼロコピーを実現する場合を説明するためのブロック図である。
【図１１】図１に示すＮＦＳサーバにおいて図９に示すファイルアクセスリクエストの境界情報に基づいてカーネル内ゼロコピーを実現する場合を説明するためのブロック図である。
【図１２】図９～図１１に示すファイルアクセスリクエストのフォーマット例を示す図である。
【図１３】図１に示すストレージシステムにおいて外部ノードからの二次記憶装置へのアクセスを許容する場合の構成を示すブロック図である。
【図１４】図１に示すストレージシステムにおいて外部ノードからのネームサーバへのアクセスを許容する場合の構成を示すブロック図である。
【図１５】図１に示すストレージシステムにおいて外部ノードからのリダイレクタ経由のアクセスと直接アクセスとを許容する場合の構成示すブロック図である。
【図１６】ネットワーク上での複数ノード（クライアント）間のファイル共有を実現する従来の手法を説明するためのブロック図である。
【符号の説明】
１　ストレージシステム
２　クライアント
３　外部ネットワーク（ギガビットイーサネット）
１１　リダイレクタ（Redirector；ファイルサーバ管理ノード）
１１ａ　ギガビットイーサネットカード
１１ｂ　インフィニバンドカード
１１ｃ　ネットワークプロセッサ
１１ｄ　メモリ（主記憶部）
１１ｅ　ＰＣＩ（Peripheral Component Interconnect））バス
１１ｆ　キャッシュメモリ（キャッシュ部）
１２－１～１２－ｎ　ＮＦＳ（Network File System）サーバ（ファイルサーバ）
１２′　キャッシュサーバ
１２ａ，１３ａ　ＣＰＵ（Central Processing Unit）
１２ｂ，１３ｂ　メモリ（主記憶部）
１２ｃ，１３ｃ　インタフェースカード（ＩＢ－ＩＦ）
１３　ネームサーバ
１４　高速（内部）ネットワーク〔インフィニバンド（Infiniband）〕スイッチ）
１５　共有メモリ（Shared Memory）
１５－１～１５－ｍ　メモリカード（Shared Memory Card）
１６　ＩＢ－ＦＣカード
１７　二次記憶装置
１８　ＦＣスイッチ
１９　外部ノード
２０　ネットワークディスクアダプタ
２１　ヘッダ部
２１ａ　物理レイヤヘッダ（Phy Header）
２１ｂ　ＩＰヘッダ（Internet Protocol Header）
２１ｃ　ＴＣＰヘッダ（Transmission Control Protocol Header）
２１ｄ　ＮＦＳヘッダ
２２　実ファイルデータ部
２３　ヘッダオフセット値（境界情報）
１１１　リクエスト解析部
１１１ａ　ヘッダオフセット値解析部
１１１ｂ　ヘッダオフセット値付加部
１１２　リクエスト転送部
１１３　転送履歴記録部
１１４　負荷監視部
１１５　ＮＦＳサーバ負荷監視デーモン（daemon）
１１６　応答部
１２１，１３１　引継ぎ情報記録部
１２２　ＮＩＣ（Network Interface Card）ドライバ（ネットワークドライバ）
１２３，１２４　バッファ（mbuf）
１２５　ファイルシステム部
１２６　ファイルシステムバッファ
１３２　ＮＦＳサーバ監視デーモン
１３３　ネームサーバ監視デーモン
１３４　異常検出部
１３５　引継ぎ指示部

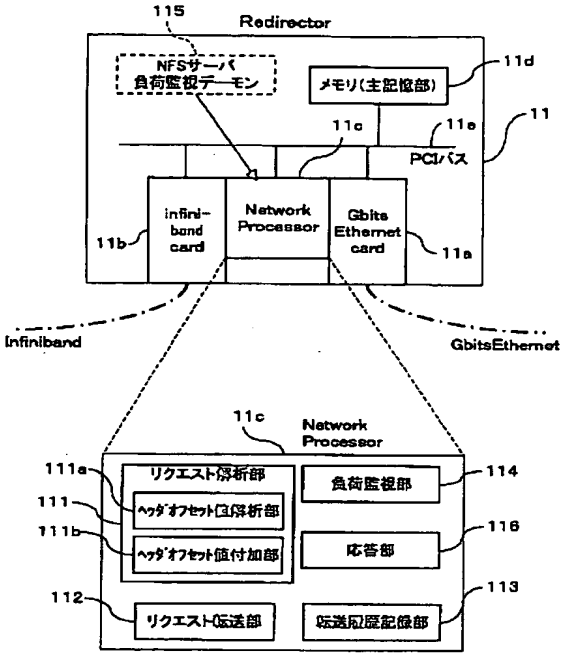
【図１２】



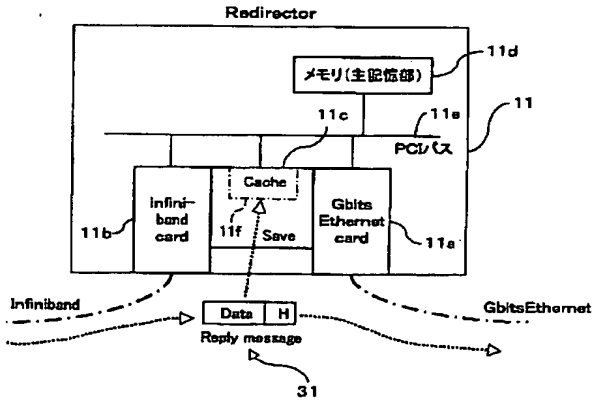
【図１】


Client
Client
Redirector
Name Server
NFS Server
NFS Server
IB-Switch
Shared Memory Card
IB-FC Card
IB-FC Card
IB-FC Card
Disk
DTL
Disk

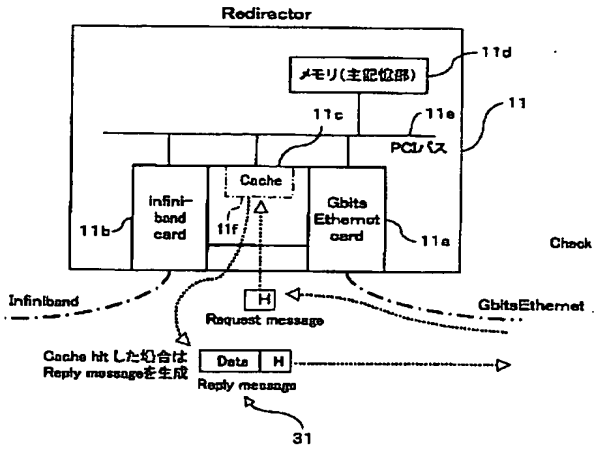
【図２】


Redirector
NFSサーバ負荷監視デーモン
メモリ(主記憶部)
PCIバス
infini-band card
Network Processor
Gbits Ethernet card
Infiniband
GbitsEthernet
Network Processor
リクエスト解析部
ヘッダオフセット値解析部
ヘッダオフセット値付加部
負荷監視部
応答部
リクエスト転送部
転送履歴記憶部

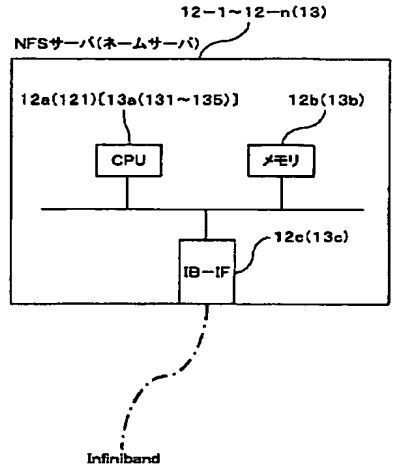
【図３】


Redirector
メモリ(主記憶部)
PCIバス
Cache
infini-band card
Save
Gbits Ethernet card
Infiniband
GbitsEthernet
Data
H
Reply message

【図４】

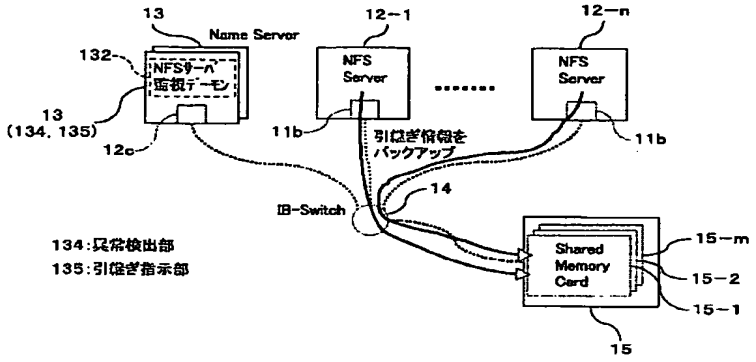

Redirector
メモリ(主記憶部)
PCIバス
Cache
infini-band card
Gbits Ethernet card
Check
Infiniband
GbitsEthernet
H
Request message
Cache hit した場合は Reply messageを生成
Data
H
Reply message

【図５】

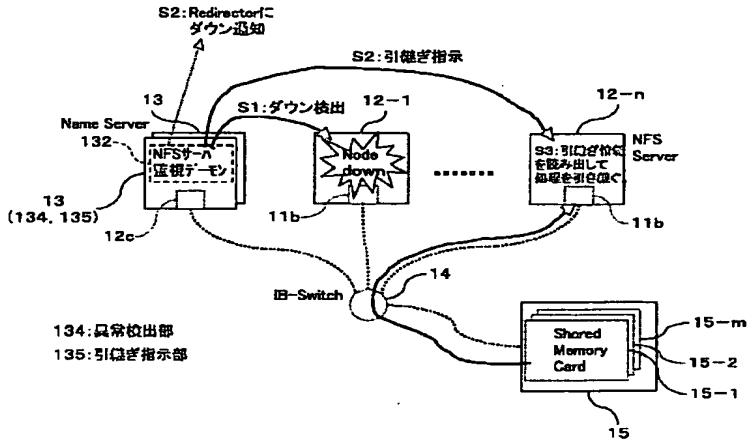


121, 131:引継ぎ情報記憶部
132:NFSサーバ監視デーモン
133:ネームサーバ監視デーモン
134:異常検出部
135:引継ぎ指示部

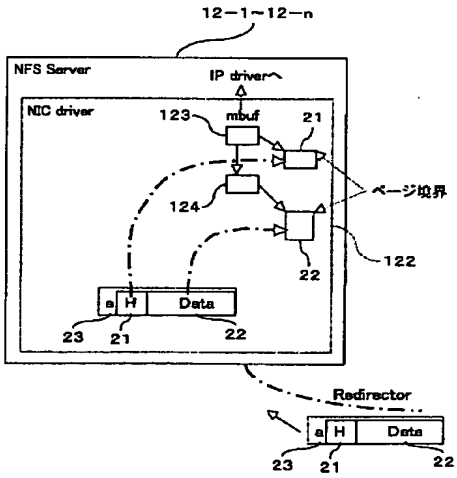
【図６】



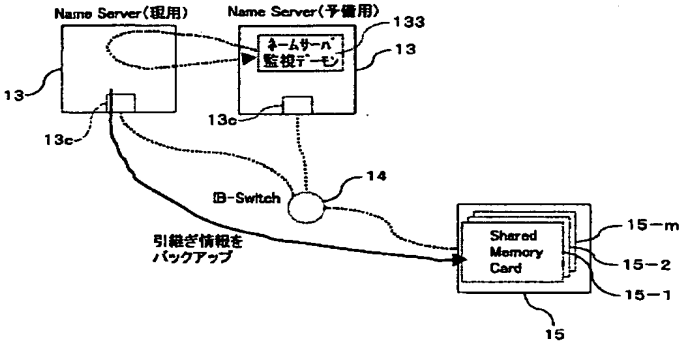
【図７】



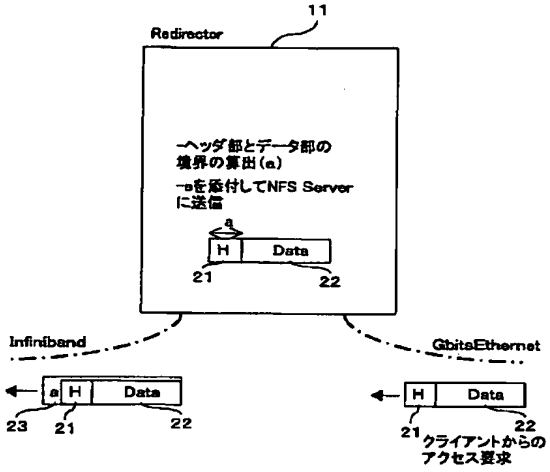
【図１０】



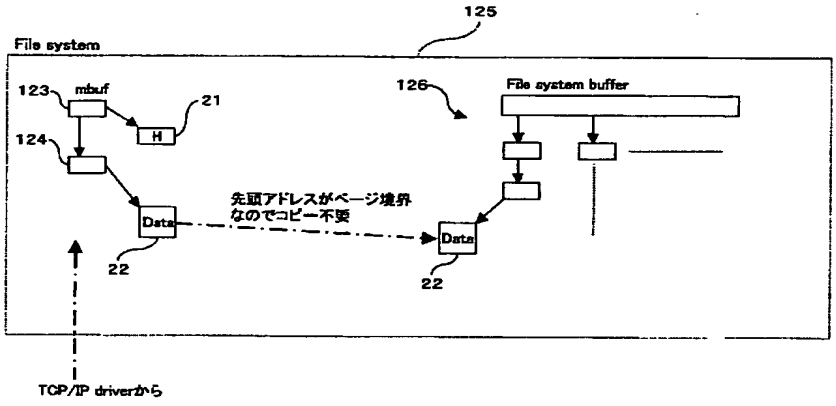
【図８】


Name Server(現用)
Name Server(予備用)
133
ネームサーバ
監視デーモン
13
13
13c
13c
14
IB-Switch
引継ぎ情報を
バックアップ
Shared
Memory
Card
15−m
15−2
15−1
15

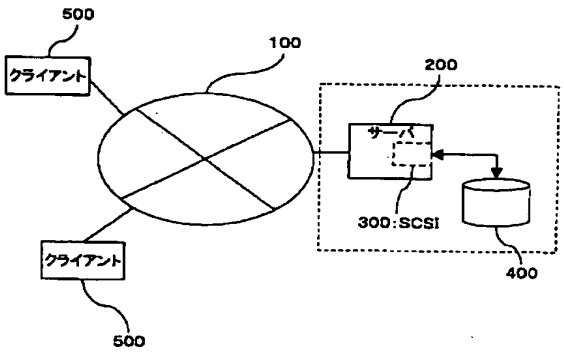
【図９】


11
Redirector
−ヘッダ部とデータ部の
境界の算出(a)
−aを添付してNFS Server
に送信
a
H
Data
21
22
Infiniband
GbitsEthernet
a
H
Data
23
21
22
H
Data
21
22
クライアントからの
アクセス要求

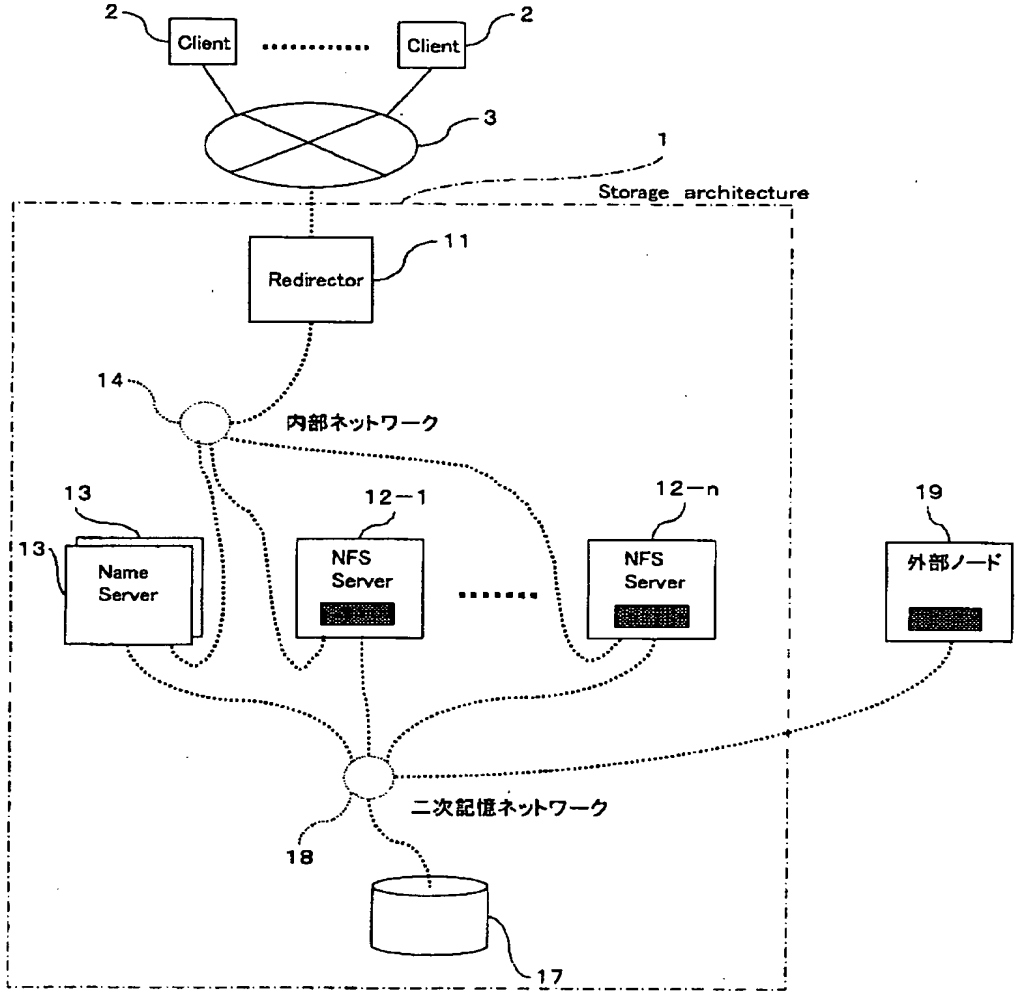
【図１１】


125
File system
123
mbuf
21
H
124
126
File system buffer
Data
22
先頭アドレスがページ境界
なのでコピー不要
Data
22
TCP/IP driverから

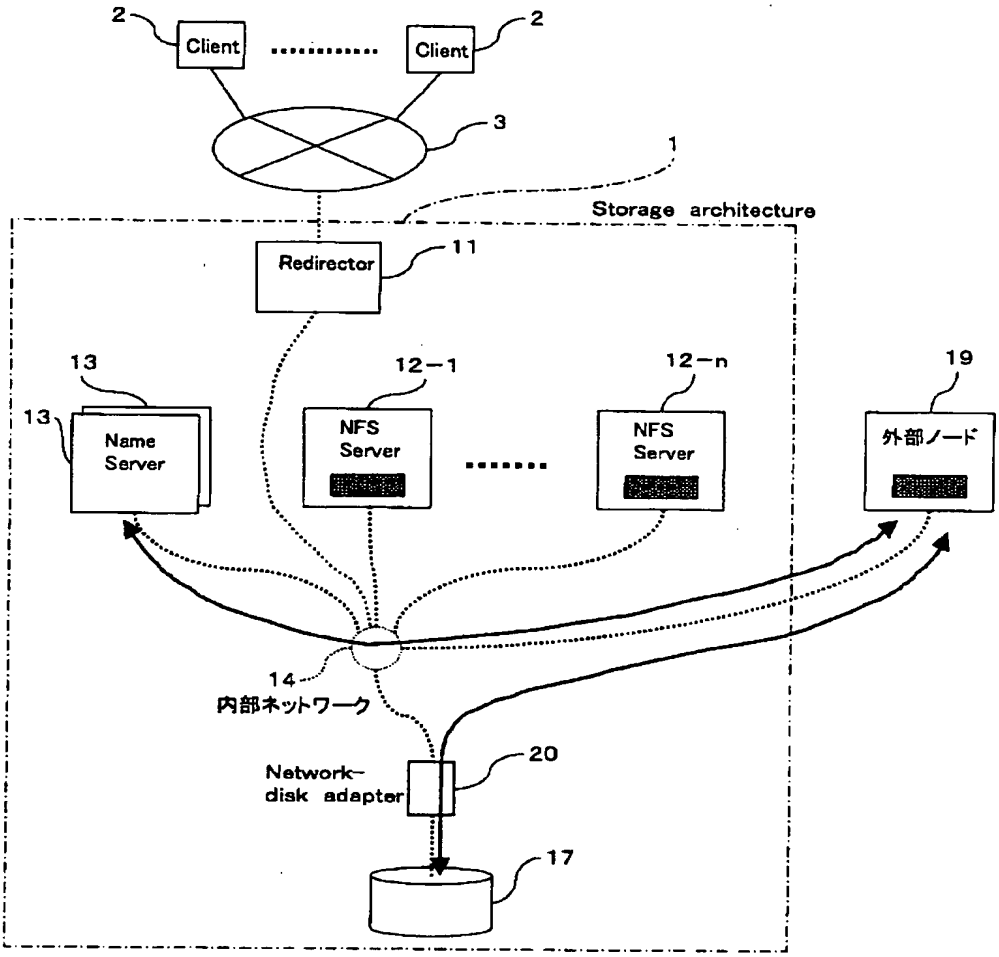
【図１６】


500
クライアント
100
200
サーバ
300:SCSI
400
クライアント
500

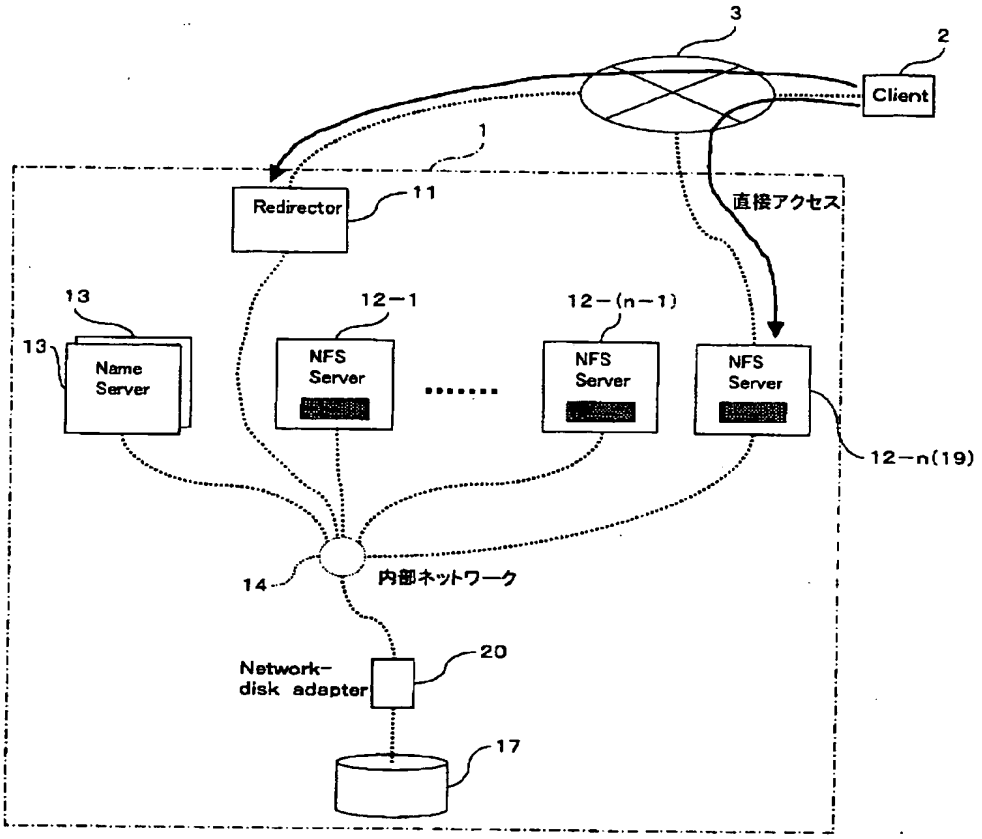
【図１３】


2
Client
Client
2
3
1
Storage architecture
Redirector
11
14
内部ネットワーク
13
12－1
12－n
19
13
Name Server
NFS Server
NFS Server
外部ノード
二次記憶ネットワーク
18
17

【図１４】


2
Client
Client
2
3
1
Storage architecture
Redirector
11
13
13
Name Server
12−1
NFS Server
12−n
NFS Server
19
外部ノード
14
内部ネットワーク
Network−disk adapter
20
17

【図１５】


3
2
Client
1
Redirector
11
直接アクセス
13
13
Name Server
12－1
NFS Server
12－(n－1)
NFS Server
NFS Server
12－n(19)
14
内部ネットワーク
Network－disk adapter
20
17